# Listening to the Noise:
# Random Fluctuations Reveal Gene Network Parameters

Brian Munsky[1,*], Brooke Trinh[2], and Mustafa Khammash[3,*]

[1] CCS-3 and CNLS,
Los Alamos National Lab,
Los Alamos, NM 87545, USA
brian.munsky@gmail.com

[2]Department of Molecular, Cellular and Developmental Biology,
University of California,
Santa Barbara, CA 93106, USA

[3] Center for Control, Dynamical Systems and Computations,
University of California,
Santa Barbara, CA 93106, USA
khammash@engr.ucsb.edu
Fax: 805-893-8651

[*] To whom all correspondence should be addressed

**Keywords:** Gene Regulatory Networks / Stochastic Biological Processes / System Identification

**Subject Categories:** Computational Methods / Simulation and data analysis

**Word count:** 3203 (22,208 characters)

## Abstract

The cellular environment is abuzz with noise attributed to random particle motion that takes part in gene expression and subsequent interactions. In this noisy environment, clonal cell populations exhibit cell-to-cell variability that can manifest significant phenotypic differences. Noise induced stochastic fluctuations in cellular constituents can be measured and their statistics quantified. We show that these random fluctuations carry within them valuable information about the underlying genetic network. Far from being a nuisance, the ever-present cellular noise acts as a rich source of excitation that, when processed through a gene network, carries its distinctive fingerprint that encodes a wealth of information about that network. We demonstrate that in some cases the analysis of these random fluctuations enables the full identification of network parameters, including those that may otherwise be difficult to measure. This establishes a potentially powerful approach for the identification of gene networks and offers a new window into the workings of these networks. **Keywords**: Gene Regulatory Networks / Stochastic Biological Processes / System Identification.

# Introduction

Computational modeling in biology seeks to reduce complex systems to their essential components and functions, thereby arriving at a deeper understanding of biological phenomena. However, measuring or estimating key model parameters can be difficult when measurement noise corrupts experimental data. Thus, when cellular variability or "noise" (Elowitz, et al, 2002) leads to measurement fluctuations, this may appear deleterious. This is not the case. Just as white noise inputs help to identify dynamical system parameters (Ljung, 1999; Cinquemani, 2009), so too can characterization of noise dynamics elucidate natural mechanisms. For example, steady state noise characteristics can distinguish between different logical structures such as AND or OR gates (Warmflash & Dinner, 2008). At the same time, temporal measurements of transient dynamics can aid in the construction of reaction pathways (Arkin et al., 1997). In combination, noise and temporal analyses yield powerful tools for parameter identification. For example, the averages of correlations in cell expression at many time points reveal feed-forward loops in the galactose metabolism genes of *E. coli* (Dunlop et al, 2008). Similarly, manipulating certain gene network transcription rates while observing the response of statistical cumulants can help to identify reaction rates for some gene regulatory networks (Rafford et al, 2008). In this paper, we examine the possibility of identifying system parameters and mechanisms directly from single cell distributions, such as those obtainable with flow cytometry, without time-varying control and at only a handful of different time points. We prove that the analysis of variability provides more information that the mean behavior alone. And we illustrate our approach's potential with numerical and experimental analyses of common gene regulatory networks.

# Results and Discussion

**Gene Expression Model**. We adopt the gene expression model in (Thattai & van Oudenaarden, 2001) characterized by random integer numbers of mRNA and protein molecules: $R$ and $P$, respectively. Transcription, translation, and degradation events change the system state by altering these numbers. mRNA changes are modeled as random events that occur according to exponentially distributed waiting times that depend on the transcription and degradation rates $k_r$ and $\gamma_r$. Thus, given a state of $r$ mRNA molecules, the probability that a single mRNA molecule is degraded within the time increment $dt$ is given by $r \cdot (\gamma_r \cdot dt)$. Similarly, translation and degradation of proteins are dictated by rates $k_p$ and $\gamma_p$. The resulting stochastic model is represented by a continuous-time, discrete-state Markov process. The probability of finding the system in a given state $(R(t) = r, P(t) = p)$ is fully characterized by the system's master equation, from which the evolution of moments $\mathbb{E}[R(t)], \mathbb{E}[P(t)], \mathbb{E}[R^2(t)], \ldots$ can be described (see S.1).

Our first finding is that *all parameters of this model are identifiable from cell population distributions of protein/mRNA measured at as few as two time instants.* In contrast, two time measurements of mRNA/protein population averages are never sufficient for identifiability. To show this, it suffices to use first and second-order moments, or equivalently means, variances, and covariances of proteins and mRNAs, instead of full distributions. At a given time, $t$, each such measurement yields a vector: $\mathbf{v}(t) = \left(\mathbb{E}[R(t)], \mathbb{E}[P(t)], \mathbb{E}[R(t)^2], \mathbb{E}[P(t)^2], \mathbb{E}[R(t)P(t)]\right)$. Given $\mathbf{v}(t_0)$ and $\mathbf{v}(t_1)$ at two distinct time instants $t_0 < t_1$, there generically exists a set of parameters $k_r, k_p, \gamma_r, \gamma_p$ that uniquely gives these measurements–all other parameter sets yield different measurements (see Figs. 1E, 2A). We illustrate this here for transcription only (S.3 provides an implicit expression for the parameters of the full model). Suppose that $\{\mu_0, \mu_1\}$ and $\{\sigma_0^2, \sigma_1^2\}$ represent the measured mRNA mean and variance at two times $t_0 < t_1 < \infty$. Then *the parameters,* $\{k_r, \gamma_r\}$ *are fully identifiable*, and

$$\gamma_r = -\frac{1}{2\tau} \log\left(\frac{\sigma_1^2 - \mu_1}{\sigma_0^2 - \mu_0}\right), \qquad k_r = \gamma_r \frac{\mu_1 - \exp(-\gamma_r \tau)\mu_0}{1 - \exp(-\gamma_r \tau)}, \qquad \text{where } \tau := t_1 - t_0.$$

Thus, the statistics, $\{\mu_0, \sigma_0^2, \mu_1, \sigma_1^2\}$, contain sufficient information to identify the model parameters. On the other hand, measurement of just the population averages, e.g. $\mathbb{E}[R]$, is insufficient for identifiability, and there exists an infinite set of parameters $\{k_r, \gamma_r\}$, that is consistent with the same two mean measurements $\mu_0$ and $\mu_1$.

While parameters are identifiable from transient moment measurements, we find that it is *impossible* to identify all parameters from *stationary moments.* Measuring means, variances, and other statistics after all the transients have died away represents a lost opportunity to peek into the cell's inner workings and to recover the network parameters. For example, two different parameter sets may produce very different protein distributions at short times (Fig. 1D) but indistinguishable distributions at long times (Fig. 1E). S.2 provides a proof that stationary moments of any arbitrary order are insufficient to uniquely identify the model parameters $k_r, k_p, \gamma_r, \gamma_p$. Such stationary distributions will only enable the determination of relative parameter values, but any positive scaling of these values would produce the exact same measurements for $\mathbf{v}_\infty$. We note that stationary correlations, e.g. $\mathbb{E}[R(t)R(t+\tau)]$ for small time steps, $\tau$, could also provide the necessary dynamical information, but taking such measurements is more difficult and requires the tracking of individual cells between measurement times.

Having determined that full identification is achievable using two measurements of all first and second order moments, we now explore the effect of partial moment measurements. We consider two new scenarios: a) only $\{\mathbb{E}[R], \mathbb{E}[P]\}$ measurements are available; and b) only $\{\mathbb{E}[P], \mathbb{E}[P^2]\}$ measurements are available. For each scenario, Fig. 2A shows the number of measurements needed for parameter identifiability and demonstrates the advantage of using full second order statistics. Furthermore, the performance with partial information depends on which partial information is used. When protein and mRNA mean measurements alone are used, full parameter identifiability is possible with three measurements. But with only protein mean and variance measurements, at least five time measurements are needed. When only protein mean measurement are available, full identifiability is impossible, regardless of the number of measurements (see S.4).

Time measurements of moment dynamics impose nonlinear algebraic constraints on model parameters. The above results can be understood by exploring how many such constraints are needed to uniquely solve for the unknown parameters. The gene expression model has $p = 4$ unknown parameters and five unknown initial conditions (moments at $t = 0$). Thus, one would expect that at least nine independent measurements are needed to identify these unknowns. The five elements of $\mathbf{v}$ at $t_0$ and $t_1$ provide ten pieces of information, and are generally sufficient (see Fig. 2A). Conversely, in a model of just the mean values $\{\mathbb{E}[R(t)], \mathbb{E}[P(t)]\}$, there are $p = 4$ parameters and two initial conditions, and one expects that at least six independent pieces of information would be needed for the identification. Indeed, at least three time measurements are required and two measurements are never enough (see Fig. 2A). However, for a model that describes only protein mean and variance measurements, at least five time measurements are needed for full parameter identifiability. In this case, the dynamics of $\{\mathbb{E}[P], \mathbb{E}[P^2]\}$ are coupled to those of $\{\mathbb{E}[R], \mathbb{E}[R^2], \mathbb{E}[RP]$, and the additional measurements are needed to identify the initial values for these. Finally, we note that in these cases, the number of measurements needed for parameter identification are far fewer than the $2p + 1$ measurements that were shown in (Sontag, 2001) to be *sufficient* for identification of the $p$ unknown parameters of a general nonlinear dynamical system.

The results above establish the principle that transient measurements of full second order moments carry information that allows one to identify all model parameters, at least assuming noise-free measurements. If the measurements are corrupted by noise, it is often possible to compensate with a larger number of measurements. To illustrate this, we have conducted 100 simulated identification studies in which the unknown parameters were taken from a broad log-normal distribution (Fig. 2B). For these, we supposed that $\mathbf{v}_j := \mathbf{v}(t_j)$ could be measured at $m$ equally separated time points $\{t_0, \ldots, t_{m-1}\}$, and that each measurement had unknown errors of $\pm 10\%$. To explore the effect of incomplete measurements, we performed the identification method for the three data scenarios considered earlier: 1) all moments; 2) only the means; and 3) only the *protein* means and variances. For each scenario, we investigated the impact on parameter identification of using an increasing number of noisy measurements obtained from a different number independent experiments (with different randomly chosen unknown initial conditions).

As more data was gathered, the effects of measurement error were overcome and the probability of successful identification increased for every strategy (see Fig. 2C). With many measurements, the parameters and the unknown initial conditions of the mRNAs and proteins could be resolved even from inaccurate protein data alone–provided that it included information on the protein variance. All of the above numerical experiments were conducted assuming that the initial conditions were unknown; if the initial conditions were known or specified, we found that the identification was even more successful (see supplemental Fig. 5). We have thus demonstrated that for the simple gene expression model, cellular noise *enhances* the opportunity for system parameter identification, whereas measurement noise impedes it. The deleterious effects of measurement noise can be overcome by increasing the number of measurements.

**Experimental Identification of *lac* Induction.** Among the most studied gene regulatory elements is the *lac* operon of *E. coli*. This mechanism has been used to construct toggle switches (Gardner et al, 2000; Kobayashi et al, 2004), genetic oscillators (Elowitz & Liebler, 2000;

Atkinson et al., 2003) and logical circuits (Weiss, 2001). Despite its ubiquitous use, precise *in vivo* single cell quantification of the system remains insufficient. Indeed, most such quantification attempts have come from *in vitro* experiments or population level studies. For example, the *lac* repressor dissociation constant has been estimated to be $K_d = 10^{-11}\mathrm{M}$ to $10^{-9}\mathrm{M}$ (Oehler et al., 1990). In an *E. coli* cell with a volume of $10^{-15}\mathrm{L}$, such dissociation constants mean the occupancy of the *lac* promoter when there are ten such molecules is 94-99.94%. At best, such measurements have only a probabilistic meaning at the level of single cells; at worst, they have no relevance at all as other mechanisms, such as non-specific binding (Kao-Huang et al., 1977), take on much greater significance.

We used flow cytometry experiments and computational analyses to identify a parameter set to describe the *in vivo* single cell dynamics of green fluorescent protein (GFP) controlled by the $lac$ operon under Isopropyl-$\beta$-D-thio-galactoside (IPTG) induction (see Fig. 3A and Methods). We explored the response of the system at several IPTG levels and at multiple time points. While many mechanistic models may capture the available data, we focused on the simplest consistent model, which consists of diffusion of IPTG into the cell, $[\mathrm{IPTG}]_{\mathrm{IN}} = [\mathrm{IPTG}]_{\mathrm{OUT}} \cdot (1 - \exp(-rt))$, and four basic reactions, $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$, and $\mathcal{R}_4$ corresponding to production and degradation of LacI and GFP.

$$\mathcal{R}_1 : \phi \xrightarrow{w_1} \text{LacI}, \quad \mathcal{R}_2 : \text{LacI} \xrightarrow{w_2} \phi \quad \mathcal{R}_3 : \phi \xrightarrow{w_3} \text{GFP}, \quad \mathcal{R}_4 : \text{GFP} \xrightarrow{w_4} \phi$$

The production of LacI is constant, $w_1 = k_L$, corresponding to constitutive expression. However, production of GFP is a nonlinear function of the LacI level:

$$w_3([\mathrm{LacI}]) = \frac{k_G}{1 + \alpha[\mathrm{LacI}]^{\eta}},$$

where $k_G$ is the unrepressed GFP production rate, $\alpha$ describes LacI occupancy strength, and the Hill coefficient, $\eta$, accounts for cooperative binding of LacI. The GFP degradation rate, $w_4 = \delta_G \cdot [\mathrm{GFP}]$, is fixed, but LacI can be degraded or inactivated by IPTG such that the total

LacI removal depends upon the IPTG concentration and is assumed to have the form $w_2 = \delta_L \cdot [\text{LacI}]$, where $\delta_L = \delta_L^{(0)} + \delta_L^{(1)}[\text{IPTG}]_{\text{IN}}$. The model also explicitly characterizes uncertainties in the flow cytometry measurements (see Methods). In total, there are ten unknown positive real parameters for the regulatory system, $\mathbf{\Lambda} = \{k_L, k_G, \delta_L^{(0)}, \delta_L^{(1)}, \delta_G, \alpha, \eta, r, \mu_{\text{GFP}}, \sigma_{\text{GFP}}^2\} \in \mathbb{R}_+^{10}$.

The measured fluorescence histograms at different times and different IPTG levels (Fig. 3) cannot adequately be captured using low order moments. Furthermore, since $w_G$ is a nonlinear function of LacI, there is no known analytical expression for the statistical moments of GFP. Instead, we used a new method, called Finite State Projection (FSP), to identify the unknown parameters based on their probability densities (see Methods). In the identification routine, a parameter search was conducted to find parameter sets such that the total predicted fluorescence distribution was as close as possible to the measured distribution in a least squares sense for all time points and IPTG levels.

Fig. 3B shows that the identified model results match the experimentally measured distributions exceptionally well. However, with the full set of ten unknowns in $\mathbf{\Lambda}$, this identification is not unique, and we found multiple parameter sets which provided equally good fits. However, by utilizing additional information about the system, we could reduce the the uncertainty of the identification. In particular, assuming that GFP is lost solely to dilution, we could specify the rate $\delta_G = 3.8 \times 10^{-4}\text{N}^{-1}s^{-1}$, corresponding to a half life of thirty minutes. The remaining nine parameters could then be identified as:

$$\left\{ \begin{array}{lll} k_L = 1.7 \times 10^{-3}\ \text{s}^{-1} & k_G = 1.0 \times 10^{-1}\ \text{s}^{-1} & \eta = 2.1 \\ \delta_L^{(0)} = 3.1 \times 10^{-4}\ \text{N}^{-1}\text{s}^{-1} & \delta_L^{(1)} = 5.0 \times 10^{-2}\ (\mu\text{M} \cdot \text{N})^{-1}\text{s}^{-1} & \alpha = 1.3 \times 10^4\ \text{N}^{-\eta} \\ r = 2.8 \times 10^{-5}\ \text{s}^{-1} & \mu_{\text{GFP}} = 220\ \text{AU} & \sigma_{\text{GFP}} = 390\ \text{AU} \end{array} \right\},$$

where N refers to molecule number.

Since the assumed model represents a simplified description of multiple events (folding dynamics, elongation, etc...), these parameters are best viewed as model-specific empirical measurements. Still, it is possible to make some comparisons between the identified parameters and previous analyses. First, the production and degradation rates of LacI yield a mean number of $k_L/\gamma_L^{(0)} \approx 5$ molecules per cell at steady state in the absence of IPTG, on the same magnitude of wild-type levels of about ten per cell. Second, the level of LacI required for half occupancy of the *lac* operon is $[\mathrm{LacI}]_{1/2} = (1/\alpha)^{1/\eta} = 0.012$ which compares well to values 0.006-0.6 molecules ($10^{-11} - 10^{-9}$ M, Oehler et al., 1990). Third, a Hill coefficient of 2.1 is reasonable considering that LacI binds to the operon as a tetramer. Finally, the degradation rate LacI, $\delta_L^{(0)}$ is close to the dilution rate of $3.8 \times 10^{-4}\mathrm{N}^{-1}s^{-1}$, reflecting the high stability of that protein. In addition to comparing the parameters to values in the literature, we have used the parameter set identified from $\{5, 10, 20\}\mu$M IPTG induction to predict the fluorescence under $\{40, 100\}\mu$M IPTG. Fig. 3C shows that these predictions match the subsequent experimental measurements very well despite the vastly different shapes observed at the high induction levels.

With single cell experimental techniques, it has become possible to efficiently measure fluctuations in cell constituents. When properly extracted and processed with rapidly improving computational tools, these measurements contain sufficiently rich information as to enable the unique identification of parameters. We have shown that transient dynamics are important to this effort, and in principle, identification can be accomplished when accurate distributions are measured at only two distinct time points. More time points are needed if the distributions are poorly measured, but the idea remains the same. We have demonstrated the potential of our approach by experimentally identifying a predictive model of *lac* regulation from flow cytometry data. Hence, the proposed integration of single cell measurements and stochastic analyses establishes a promising approach that offers new windows into the workings of cellular networks.

# Methods

**Media and Reagents.** Cells were grown in Luria-Bertani (LB) 1% tryptone, 0.5% yeast extract, 0.4% NaCl containing Isopropyl B-D-thiogalactoside (IPTG) at the concentrations noted. To select for plasmid maintenance, antibiotics were used at the following concentrations: 100$\mu$g/ml ampicillin (amp); 40$\mu$g/ml kanamycin (kan); 12.5 $\mu$g/ml tetracycline (tet).

**Bacterial Strains and Plasmids.** The *E. coli* strain used was DL5905: *E. coli* K-12 (isolate MC4100) containing *[F' proAB lacI$_q$Z$\Delta$M15 Tn10 (Tet$^r$)]* from strain XL-1 Blue (Stratagene) and plasmid pDAL812. To construct plasmid pDAL812, GFP(LVA) (Anderdson et al., 1998), was PCR amplified from plasmid pRK9 (a gift from John Cronan) using the forward primer (5'-CAA CAA AGA TCT ATT AAA GAG GAG AAA TTA AGC ATG AGT AAA GGA GAA GAA CTT TTC A-3') which includes a BglII site and removes an SphI site from the original pRK9 sequence, and the reverse primer (5'-CAA CAA GCA TGC ATT AAG CTA CTA AAG CGT AGT TTT CGT CGT TTG C-3') which adds an SphI site. This fragment was digested with BglII and SphI and cloned into the BglII and SphI sites of pLAC33 (Warren et al., 2000), removing a portion of the Tet$^R$ cassette.

**Fluorescence Induction Experiments.** Twenty-four separate cell cultures were allowed to grow in LB broth containing the appropriate antibiotics to an approximate OD600 of 0.2 and were then induced with $\{0, 5, 10, 20, 40, 100\}\mu$M concentrations of IPTG at the times of 5, 4, 3, and 0 hours before flow cytometry measurements. Flow cytometry was carried out using a BD Biosciences FACSAria instrument with a 100$\mu$m sorting nozzle at low pressure. GFP(LVA) was excited using a 488nm blue laser and detected using 530/30nm filter. For each sample, 1,000,000 events were collected. To ensure repeatability, the experiments were conducted twice each on a separate days.

**GFP Induction Model**. The stochastic model for the IPTG-GFP induction is composed of the four non-linear production / degradation reactions given in the main text. The rates of these reactions depend upon the integer populations of the proteins LacI and GFP as well as the set of non-negative parameters, $\{k_L, k_G, \delta_L^{(0,1)}, \delta_G, \alpha, r, \eta\} \in \mathbb{R}^8$. For the stochastic system modeled here, the joint (LacI, GFP) probability distributions of both proteins evolve according to the infinite dimensional Chemical Master Equation (vanKampen, 2001). This can in turn be expressed as an infinite set of linear ordinary differential equations, $\dot{\mathbf{P}}(t, \mathbf{\Lambda}) = \mathbf{A}(t, \mathbf{\Lambda}) \cdot \mathbf{P}(t, \mathbf{\Lambda})$. Unlike in the simple transcription/translation model, the toggle reactions are non-linear, and the CME has no known exact solution. We use a finite state projection approach (Munsky & Khammash, 2006) that makes it possible to approximate the solution to any degree of accuracy. For any error tolerance $\varepsilon > 0$, we systematically find a finite-dimensional projected system $\dot{\mathbf{P}}^{FSP}(t, \mathbf{\Lambda}) = \mathbf{A}_J(t, \mathbf{\Lambda}) \cdot \mathbf{P}^{FSP}(t, \mathbf{\Lambda})$ whose solution is within the desired tolerance. More precisely:

$$\left\| \begin{bmatrix} \mathbf{P}_J(t, \mathbf{\Lambda}) \\ \mathbf{P}_{J'}(t, \mathbf{\Lambda}) \end{bmatrix} - \begin{bmatrix} \mathbf{P}^{FSP}(t, \mathbf{\Lambda}) \\ \mathbf{0} \end{bmatrix} \right\|_1 \leq \varepsilon, \text{ and } \mathbf{P}^{FSP}(0, \mathbf{\Lambda}) = \mathbf{P}_J(0, \mathbf{\Lambda}),$$

where the index vector $J$ denotes the set of states included in the projection, $\mathbf{P}_J$ is the corresponding probabilities of those states, and $\mathbf{A}_J$ is the corresponding principle submatrix of $\mathbf{A}$ (Munsky & Khammash, 2006). The one-norm measure is used to ensure that absolute sum of the probability density error is guaranteed to lie within the tolerance. The solution of each projected master equation is found using the stiff ode solver *ode23s* in MathWorks Matlab.

**Modeling Flow Cytometry Data.** In addition to modeling the regulatory dynamics of the system, one must also account for the inherent uncertainty within measured levels of fluorescence activity. The process used to account for this uncertainty has three components. First, in an effort remove outliers in cell volume and density and thereby reduce the effects of unmodeled dynamics, each cell population was gated separately using forward and side side scatter data. Specifically, the forward and side scatter measurements were used to form a two-dimensional joint histogram with 50 × 50 logarithmically distributed bins (see Supplemental Fig. 6). The maximum point in this histogram was recorded and then the gating region was chosen to include every bin which had at least one third as many counts as the maximal bin. Second, flow cytometry measurements in the absence of IPTG have been used to calibrate for the background fluorescence of the cell populations at the various instances in time, and it has been assumed that the background fluorescence distribution, $f_{\mathrm{BG}}(x)$, is independent of the levels of IPTG, LacI and GFP. Third, each GFP molecule is assumed to emit a normally distributed random amount of fluorescence with unknown mean, $\mu_{\mathrm{GFP}}$, and variance, $\sigma^2_{\mathrm{GFP}}$, both of which are to be identified. Thus, if $p_n = p_n(t, \mathbf{\Lambda}, [\mathrm{IPTG}])$ denotes the probability of having exactly $n = \{0, 1, 2, \ldots\}$ molecules of GFP, then the probability density of having exactly $x$ arbitrary units of fluorescence due to GFP is computed as:

$$f_{\mathrm{GFP}}(x) = \sum_{n=0}^{\infty} p_n \cdot \frac{1}{\sqrt{2n\pi \cdot \sigma^2_{\mathrm{GFP}}}} \exp\left( -\frac{(x - n \cdot \mu_{\mathrm{GFP}})^2}{2n \cdot \sigma^2_{\mathrm{GFP}}} \right).$$

Finally, the total observable fluorescence is the sum of the GFP florescence plus the background noise, and the distribution of total fluorescence is found via the convolution:

$$f_{\mathrm{Tot}}(x) = \int_{-\infty}^{x} f_{\mathrm{GFP}}(x-s) \cdot f_{\mathrm{BG}}(s) \cdot ds \approx \int_{0}^{x} f_{\mathrm{GFP}}(x-s) \cdot f_{\mathrm{BG}}(s) \cdot ds.$$

**Identification Procedure.** With the FSP solution and the computation of the expected fluorescence, the identification procedure is carried out by finding the parameter vector $\mathbf{\Lambda}^\star$ that minimizes the one norm difference between the experimentally measured distribution $f^{(i)}_{\mathrm{Meas}}(t, [\mathrm{IPTG}])$ and the numerical solution of that distribution:

$$\mathbf{\Lambda}^\star := \mathrm{argmin}_{\mathbf{\Lambda}} \left\{ \sum_i q_i \cdot \left\| f^{(i)}_{\mathrm{Meas}} - f^{(i)}_{\mathrm{Tot}} \right\|_1 \right\},$$

where the summation is taken over all of the different experimental conditions of different induction times and IPTG levels, and the weight $q_i$ specifies a relative importance to each of these measurements. These weights have been chosen such that each IPTG level has the same total importance and so that greater importance is placed upon measurements that differ most from the background fluorescence. The values for these weights are given in Fig. 3. The parameter identification is accomplished by starting with an initial parameter guess, $\mathbf{\Lambda}_0$, and then this set is updated iteratively using gradient-based and simulated annealing searches until the computed distribution matches the experimental distribution as closely as possible. The optimization procedure is repeated for multiple, randomly generated initial parameter guesses. An optimal parameter set is regarded as unique if the given solution yields the smallest achieved value for the objective function and if that parameter has been achieved during many such identification runs each beginning with different parameter guesses.

## Acknowledgements

The authors thank David Low for advice, assistance, and materials in the experimental aspects of this work. Authors acknowledge funding from the LANL LDRD program and NSF through Grants ECCS-0835847 and ECCS-0802008 and from ICB Grant DAAD19-03-D-0004 from the US Army Research Office.

# References

[1] Arkin A, Shen, P, Ross J (1997) A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements *Science* **227**:1275-1279

[2] Atkinson M, Savageau M, Myers J, Ninfa, A (2003) Development of genetic circuitry exhibiting toggle switch or oscillatory behaviour in E. coli. *Cell* **113**:597–607

[3] Cinquemani E, Milias-Argeitis A, Summers S, Lygeros J (2009) *Local Identification of Piecewise Deterministic Models of Genetic Networks, Lecture Notes in Computer Science*, Springer, **5469**:105-119

[4] Dunlop M, Cox III R, Levine J, Murray R, Elowitz M (2008) Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genetics* **40**:1493–1498

[5] Elowitz M, Levine A, Siggia E, Swain P (2002) Stochastic gene expression in a single cell. *Science* **297**:1183–1186

[6] Kao-Huang, Y, et al. (1977) Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: Measurement of DNA-bound E. coli lac repressor in vivo. *Proc Natl Acad Sci* **74**:4228–4232

[7] Ljung L (1999) *System Identification, Theory for the User.* Prentice-Hall, Inc. Upper Saddle River, NJ, USA

[8] Munsky B, Khammash M (2006) The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.* **124**:044104

[9] Ozbudak E, Thattai M, Kurtser I, Grossman A, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. *Nature Genetics* **31**:69–73

[10] Sontag E (2002) For Differential Equations with r Parameters, 2r+1 Experiments Are Enough for Identification. *J Nonlinear Sci* **12**:553–583

[11] Thattai M, van Oudenaarden A (2001) Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci* **98**:8614–8619

[12] Raffard R, Lipan O, Wong W, Tomlin C (2008) Optimal discovery of a stochastic genetic network. *Proc. 2008 Amer. Contr. Conf.*

[13] van Kampen N (2001) *Stochastic Processes in Physics and Chemistry*. Elsevier, 2001.

[14] Warmflash A, Dinner A, (2008) Signatures of combinatorial regulation in intrinsic biological noise. *Proc. Nat. Acad. Sci. USA* **105**:17262-17267.

[15] Weiss R, (2001) *Cellular Computation and Communications using Engineered Genetic Regular Networks* PhD thesis, MIT, 2001.

# Figure Legends

**Figure 1.** **(A)** Simple gene expression model representing gene transcription and translation. **(B,C)** Simulations of mRNA (green) and protein (blue) populations. The solid red lines denote the mean values, and the dashed lines are one standard deviation above and below that mean. **(D,E)** mRNA (green) and protein (blue) distributions at (D) $t = 5000s$ and (E) $t = 1000s$ for two different parameter sets but the same initial conditions.

**Figure 2.** Comparison of strategies for the identification of the gene expression model. (A) Minimum number of measurements needed for full parameter identification. (B) The log-normally distributed parameters of 100 simulated models, which combined with an initial distribution at time $t = 0$ defined the moment trajectories. (C) Percent identification success rates (within $5\%$ for all parameters) for different identification strategies, assuming that measurements had unknown errors of $\pm 10\%$ and were taken every 100 seconds.

**Figure 3.** Experimental identification of a simple construct (A) in which IPTG induces the production of GFP. (B) Experimentally measured histograms of *gfp* expression on two different days (solid blue and green lines–in arbitrary units) and the best determined parameter fit (red-dashed lines). Here each column corresponds to a different measurement time (0,3,4,5)hr after induction and each row corresponds to a different level of extra-cellular IPTG induction (5,10,20)$\mu$M. In the parameter fits, different weights were applied to each experimental condition, shown as the values $\{q\}$ in the histograms. (C) Predicted (red) then measured (blue and green) Fluorescence at (40,100)$\mu$M.

# Figures

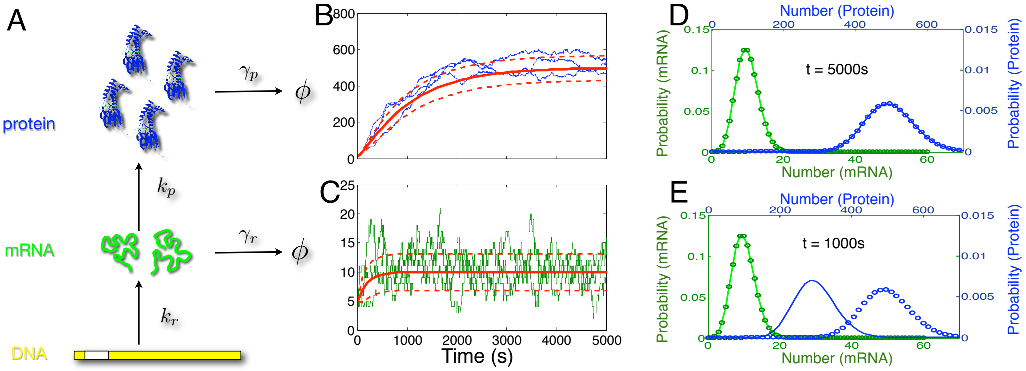
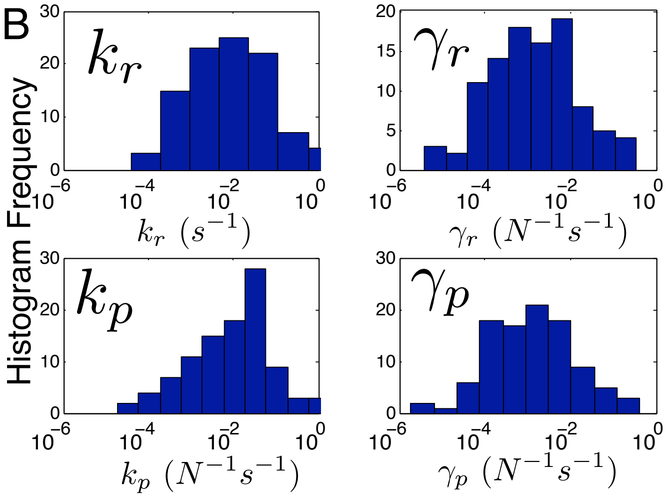


Figure 1:

A

Parameter Identification with
Noise-free Measurements (One Experiment)

| Measured Variables | Required No. of time measurements |
|---|---|
| $\mathbb{E}[R], \mathbb{E}[R^2], \mathbb{E}[P], \mathbb{E}[P^2], \mathbb{E}[RP]$ | 2 |
| $\mathbb{E}[R], \mathbb{E}[P]$ | 3 |
| $\mathbb{E}[P], \mathbb{E}[P^2]$ | 5 |



C

Percentage of Parameters Identified
with 10% Measurements Noise

| Measured Variables | No. of time measurements | No. of experiments | | |
|---|---|---|---|---|
| | | **1** | **2** | **4** |
| $\mathbb{E}[R], \mathbb{E}[R^2], \mathbb{E}[P],$<br>$\mathbb{E}[P^2], \mathbb{E}[RP]$ | **10** | 12 | 32 | 37 |
| | **20** | 22 | 39 | 60 |
| | **40** | 45 | 70 | 84 |
| $\mathbb{E}[R], \mathbb{E}[P]$ | **10** | 2 | 13 | 23 |
| | **20** | 13 | 24 | 30 |
| | **40** | 20 | 44 | 60 |
| $\mathbb{E}[P], \mathbb{E}[P^2]$ | **10** | 0 | 0 | 0 |
| | **20** | 1 | 1 | 3 |
| | **40** | 1 | 3 | 8 |

Figure 2:

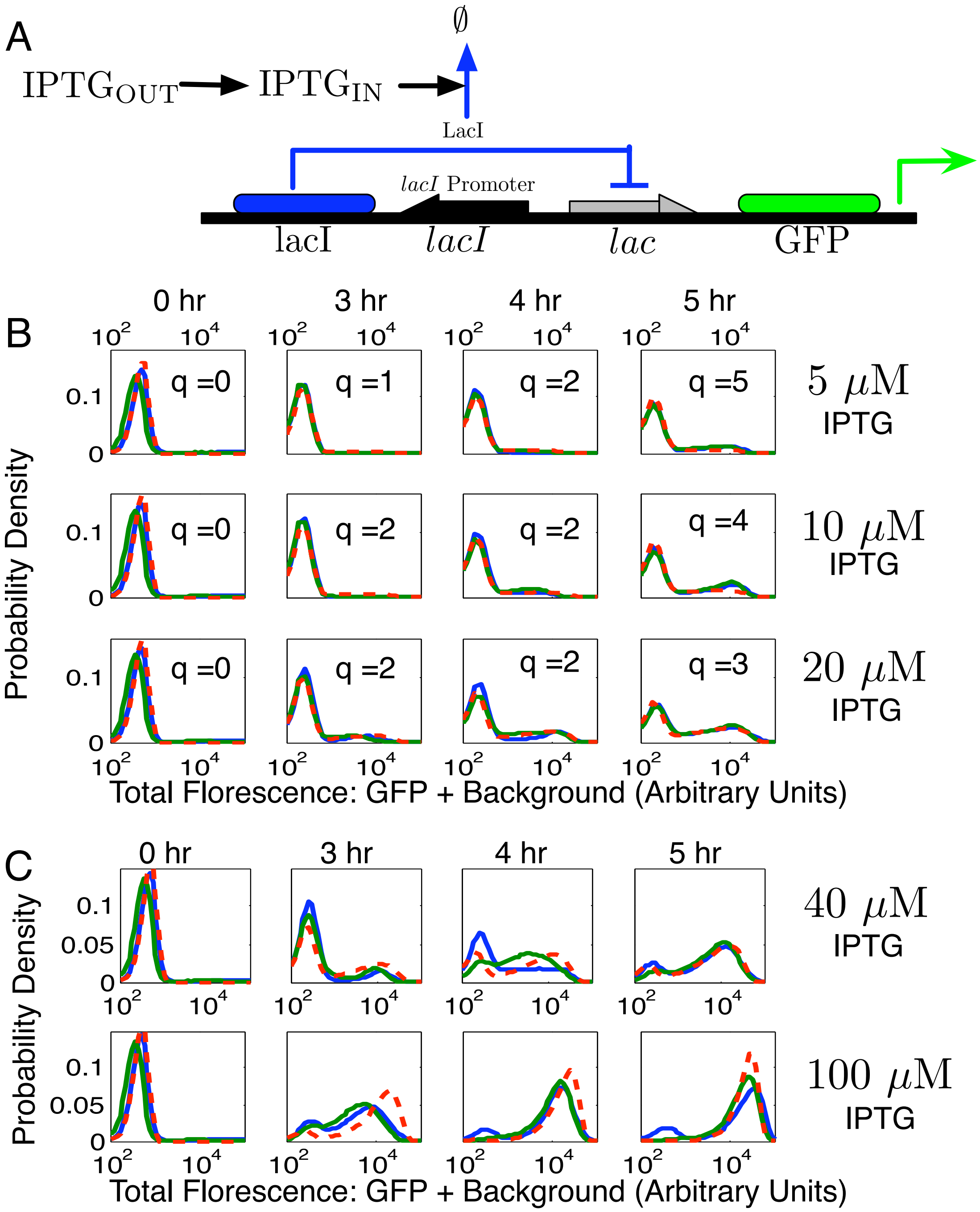

A
∅
IPTG$_{OUT}$
IPTG$_{IN}$
LacI
*lacI* Promoter
lacI
*lacI*
*lac*
GFP
B
0 hr
3 hr
4 hr
5 hr
q =0
q =1
q =2
q =5
q =0
q =2
q =2
q =4
q =0
q =2
q =2
q =3
5 $\mu$M IPTG
10 $\mu$M IPTG
20 $\mu$M IPTG
Probability Density
0.1
0
$10^2$
$10^4$
Total Florescence: GFP + Background (Arbitrary Units)
C
0 hr
3 hr
4 hr
5 hr
40 $\mu$M IPTG
100 $\mu$M IPTG
Probability Density
0.1
0.05
0
$10^2$
$10^4$
Total Florescence: GFP + Background (Arbitrary Units)

Figure 3:

# Supplemental Material

## S.1. Implicit expression for transcription and translation.

In this supplemental section, we derive explicit expressions for the evolution of the first two moments in the simple gene transcription and translation process. For this derivation, let $R$ denote the population of mRNA molecules, and let $P$ denote the population of proteins in the system. As above, these populations change through four reactions:

$$\emptyset \rightarrow R$$

$$R \rightarrow \emptyset$$

$$R \rightarrow R + P$$

$$P \rightarrow \emptyset$$

for which the propensity functions (or stochastic reaction rates) are

$$w_1 = k_r + k_{21}P,$$

$$w_2 = \gamma_r R,$$

$$w_3 = k_p R, \text{ and}$$

$$w_2 = \gamma_p P.$$

Here the term $k_{21}$ corresponds to a feedback effect that the protein is assumed to have on the transcription process. In positive feedback, $k_{21} > 0$, the protein increases transcription; in negative feedback, $k_{21} < 0$, the protein inhibits transcription. For the results in the main text, this feedback term has been set to zero.

The master equation [13] for this system can be written:

$$\begin{aligned}\dot{P}_{i,j}(t) = &-(k_r{+}k_{21}j + \gamma_r i + k_p i + \gamma_p j)P_{i,j}(t)\\ &+ (k{+}k_{21}j)P_{i-1,j}(t)\\ &+ \gamma(i+1)P_{i+1,j}(t)\\ &+ k_p i P_{i,j-1}(t)\\ &+ \gamma_p(j+1)P_{i,j+1}(t),\end{aligned} \tag{1}$$

where $P_{i,j}(t)$ is the probability that $(R, P) = (i, j)$ at the time $t$, conditioned on some initial probability distribution $\mathbf{P}(t_0)$. In this expression, the first negative term corresponds to the probability of transitions that begin at the state $(R, P) = (i, j)$ and leave to another state, and the remaining positive terms correspond to the reactions that begin at some other state $(R, P) \neq (i, j)$ and transition into the state $(i, j)$.

The mean populations of mRNA and protein molecules can be written as:

$$\begin{aligned} v_1(t) &= \mathbb{E}\{R\} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} i P_{i,j}(t)\\ v_2(t) &= \mathbb{E}\{P\} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} j P_{i,j}(t). \end{aligned} \tag{2}$$

The derivatives of these mean values are found simply by substituting (1) into (2):

$$\dot{v}_1(t) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} i\dot{P}_{i,j}(t) = k_r{+}k_{21}v_2 - \gamma_r v_1,$$

and

$$\dot{v}_2 = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} j\dot{P}_{i,j}(t) = k_p v_1 - \gamma_p v_2.$$

Similarly, expressions for the second uncentered moments can be written:

$$\begin{aligned} v_3 &= \mathbb{E}\{RR\} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} iiP_{i,j}, \\ v_4 &= \mathbb{E}\{PP\} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} jjP_{i,j}, \\ v_5 &= \mathbb{E}\{RP\} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} ijP_{i,j}, \end{aligned} \tag{3}$$

and evolve according to the set of ordinary differential equations:

$$\begin{aligned} \dot{v}_3 &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} i^2\dot{P}_{i,j}(t) \\ &= k_r + (2k_r + \gamma_r)v_1 - 2\gamma_r v_3 {+} k_{21}v_2 + 2k_{21}v_5, \\ \dot{v}_4 &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} j^2\dot{P}_{i,j} \\ &= k_p v_1 + \gamma_p v_2 - 2\gamma_p v_4, +2k_p v_5, \\ \dot{v}_5 &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} ij\dot{P}_{i,j} \\ &= k_p v_3 + k_r v_2 {+} k_{21}v_4 - (\gamma_r + \gamma_p)v_5. \end{aligned}$$

Altogether the various components of the first two moments,

$$\mathbf{v}(t) := \left[ \begin{array}{ccccc} \mathbb{E}\{R\} & \mathbb{E}\{P\} & \mathbb{E}\{RR\} & \mathbb{E}\{PP\} & \mathbb{E}\{RP\} \end{array} \right]^T,$$

evolve according to the linear time invariant ODE:

$$\begin{aligned}\dot{\mathbf{v}} &= \begin{bmatrix} -\gamma_r & k_{21} & 0 & 0 & 0 \\ k_p & -\gamma_p & 0 & 0 & 0 \\ \gamma_r + 2k_r & k_{21} & -2\gamma_r & 0 & 2k_{21} \\ k_p & \gamma_p & 0 & -2\gamma_p & 2k_p \\ 0 & k_r & k_p & k_{21} & -\gamma_r - \gamma_p \end{bmatrix} \mathbf{v} + \begin{bmatrix} k_r \\ 0 \\ k_r \\ 0 \\ 0 \end{bmatrix} \\ &= \mathbf{A}\mathbf{v} + \mathbf{b} \end{aligned} \tag{4}$$

These expressions now fully characterize the dynamics of the first two moments of mRNA and protein molecules. With these expressions one can now begin to identify the various parameters: $[k_r, \gamma_r, k_p, \gamma_p, k_{21}]$ from properly chosen experimental data sets.

### S.2. Non-Identifiability from Stationary Distributions

In this supplemental section, we show conclusively that the parameters of the transcription/translation model cannot be identified from invariant distributions alone. Suppose that the moments of the probability distribution described in (4) has an invariant distribution:

$$\mathbf{v}_\infty = \lim_{t\to\infty} [v_1, v_2, v_3, v_4, v_5]^T.$$

These steady state moments must satisfy the expression:

$$\mathbf{A}\mathbf{v}_\infty - \mathbf{b} = \mathbf{0}, \tag{5}$$

which can be rewritten in terms of the unknown parameters as:

$$\mathbf{\Psi}_\infty \mathbf{\Lambda} = \lim_{t\to\infty} \mathbf{\Psi}(t)\mathbf{\Lambda} = \mathbf{0},$$

where

$$\mathbf{\Psi}(t) = \begin{bmatrix} 1 & -v_1 & 0 & 0 & v_2 \\ 1{+}2v_1 & v_1 - 2v_3 & 0 & 0 & v_2{+}2v_5 \\ 0 & 0 & v_1 & -v_2 & 0 \\ 0 & 0 & v_1{+}2v_5 & v_2{-}2v_4 & 0 \\ v_2 & -v_5 & v_3 & -v_5 & v_4 \end{bmatrix}.$$

In Eqn. (5) there are two possible cases: (1) the rank of the matrix is full and we are left with the trivial solution $\mathbf{\Lambda} = \mathbf{0}$, or (2) the matrix has a null-space spanned by $\{\phi_1, \ldots, \phi_p\}$ and there are an infinite number of parameter sets that will result in the same invariant distribution:

$$\mathbf{\Lambda} = \sum_{i=1}^{p} \alpha_i \phi_i, \text{ for any } [\alpha_1, \ldots, \alpha_p] \in \mathbb{R}^p.$$

So long as the parameters enter linearly into the propensity functions $w(\mathbf{x}) = \sum_{\mu=1}^{M} c_\mu f(\mathbf{x})$, then one can extend this argument for any finite number of $n$ moments of the stationary distribution. This tells us that *the steady state distribution cannot provide enough information* to uniquely identify the set of system parameters. Additional information is needed. For example, if the rank of the null space is one, then the knowledge of any one parameter from the set $\mathbf{\Lambda}$ can provide an additional linearly independent equation, and can enable the unique determination of the parameters. If the rank of the null space is $p$, then at least $p$ additional, linearly independent, pieces of information will be required.

### S.3. Implicit Expressions for the Identification of Transcription and Translation Parameters from Transient Data

In this supplemental section, we show how one can obtain an implicit analytical expression for transcription and translation parameters in the absence of feedback ($k_{12} = 0$). For this we define the following variables:

$$\begin{bmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \\ z_4(t) \end{bmatrix} = \begin{bmatrix} \mu_r \\ \sigma_{rr} - \mu_r \\ \mu_p \\ \sigma_{rp} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_3 - v_1^2 - v_1 \\ v_2 \\ v_5 - v_1 v_2 \end{bmatrix}.$$

These can be shown to evolve according to the linear ODE:

$$\frac{d}{dt}\begin{bmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \\ z_4(t) \end{bmatrix} = \begin{bmatrix} -\gamma_r & 0 & 0 & 0 \\ 0 & -2\gamma_r & 0 & 0 \\ k_p & 0 & -\gamma_p & 0 \\ k_p & k_p & 0 & -(\gamma_r+\gamma_p) \end{bmatrix}\begin{bmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \\ z_4(t) \end{bmatrix} + \begin{bmatrix} k_r \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The first two equations yield $k_r$ and $\gamma_r$ as discussed above. With these, one can solve for the $z_1(t)$ and $z_2(t)$:

$$\begin{aligned} z_1(t) &= \mathrm{e}^{-\gamma_r(t-t_1)} z_1(t_1) + \frac{k_r}{\gamma_r}\left(1 - \mathrm{e}^{-\gamma_r(t-t_1)}\right) \\ z_2(t) &= \mathrm{e}^{-2\gamma_r(t-t_1)} z_2(t_1), \end{aligned}$$

and plug these expressions into the third and fourth equations. This gives the following expressions for the solution:

$$\begin{bmatrix} z_3(t_2) \\ z_4(t_2) \end{bmatrix} = \begin{bmatrix} \mathrm{e}^{-\gamma_p(t_2-t_1)} z_3(t_1) + k_p \int_{t_1}^{t_2} \mathrm{e}^{-\gamma_p(t_2-\tau)} z_1(\tau) d\tau \\ \mathrm{e}^{-(\gamma_r+\gamma_p)(t_2-t_1)} z_4(t_1) + k_p \int_{t_1}^{t_2} \mathrm{e}^{-(\gamma_r+\gamma_p)(t_2-\tau)} (z_1(\tau) + z_2(\tau)) d\tau. \end{bmatrix}.$$

One can combine many of the known quantities to gather a simpler expression

$$\begin{bmatrix} z_3(t_2) \\ z_4(t_2) \end{bmatrix} = \begin{bmatrix} \mathrm{e}^{-\gamma_p(t_2-t_1)} z_3(t_1) + k_p \int_{t_1}^{t_2} \mathrm{e}^{-\gamma_p(t_2-\tau)} z_1(\tau) d\tau \\ \mathrm{e}^{-\gamma_p(t_2-t_1)} C_1 + k_p \int_{t_1}^{t_2} \mathrm{e}^{-\gamma_p(t_2-\tau)} C_2(\tau) d\tau. \end{bmatrix}, \tag{6}$$

where

$$C_1 = \mathrm{e}^{-\gamma_r(t_2-t_1)} z_4(t_1), \text{ and}$$

$$C_2(\tau) = \mathrm{e}^{-\gamma_r(t_2-\tau)}(z_1(\tau) + z_2(\tau))$$

are known expressions. Solving the first expression in terms of $k_p$ and substituting that expression into the second yields the implicit expression for $\gamma_p$:

$$z_4(t_2) = \mathrm{e}^{-\gamma_p(t_2-t_1)} C_1 + \frac{z_3(t_2) - \mathrm{e}^{-\gamma_p(t_2-t_1)} z_3(t_1)}{\int_{t_1}^{t_2} \mathrm{e}^{-\gamma_p(t_2-\tau)} z_1(\tau) d\tau} \int_{t_1}^{t_2} \mathrm{e}^{-\gamma_p(t_2-\tau)} C_2(\tau) d\tau. \tag{7}$$

An explicit expression for $\gamma_p$ does not appear to be immediately obvious. However, by substituting in the known expressions for $C_1$, $C_2(\tau)$, and $z_3(t_1)$, one can easily plot the the left hand side of this expression as a function of $\gamma_p$. For example, consider the system with the parameter set:

$$\mathbf{\Lambda} = \begin{bmatrix} k_r \\ \gamma_r \\ k_p \\ \gamma_p \end{bmatrix} = \begin{bmatrix} 0.05 \\ 0.005 \\ 0.05 \\ 0.001 \end{bmatrix},$$

and the initial condition at $t_1 = 0$ of

$$\mathbf{z}(t_1) = \begin{bmatrix} z_1(t_1) \\ z_2(t_1) \\ z_3(t_1) \\ z_4(t_1) \end{bmatrix} = \begin{bmatrix} 1 \\ 0.2 \\ 10 \\ 5 \end{bmatrix}.$$

The corresponding response at $t_2 = 100s$ is

$$\mathbf{z}(t_2) = \begin{bmatrix} z_1(t_2) \\ z_2(t_2) \\ z_3(t_2) \\ z_4(t_2) \end{bmatrix} = \begin{bmatrix} 4.541 \\ 0.07358 \\ 23.07 \\ 14.82 \end{bmatrix}.$$

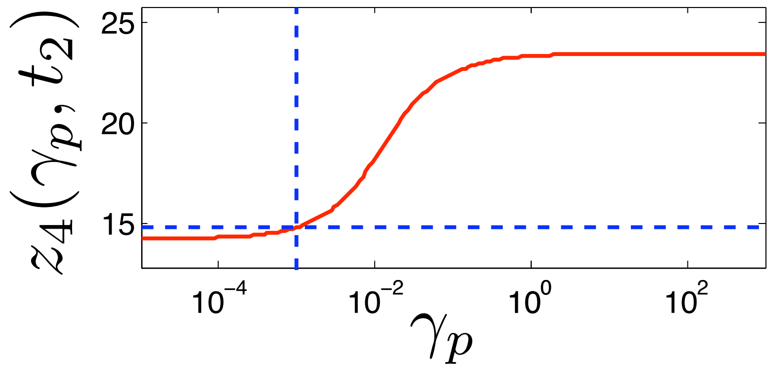


Figure 4: Implicit determination of the protein degradation rate. The red curve corresponds to the right hand side of Eqn. 7 versus $\gamma_p$. The horizontal dashed line corresponds to the measured value of $z_4(t_2)$. From the figure, one can correctly determine that $\gamma_p = 0.001$ (vertical dashed line).

Assuming that the quantities $\mathbf{z}(t_1)$ and $\mathbf{z}(t_2)$ are known exactly, then it is relatively easy to identify the first two parameters $k_r$ and $k_p$ and substitute these into the expression (7). This expression can then be plotted as a function of the unknown $\gamma_p$ as shown in Figure 4. The value of $\gamma_p$ is the value at which the expression crosses the measured value for $z_4(t_2)$, which can be found using a simple line search. From the figure it is obvious that this intersection does indeed correspond to the correct value of $\gamma_p = 0.001$. Once $k_r$, $\gamma_r$, and $\gamma_p$ are all known, it is simple to solve for $k_p$ using (6).

### S.4. Non-Identifiability from Protein Mean alone

In this section we show analytically that the parameters of the transcription/translation cannot be identified from the protein mean alone. Consider the expressions for the mRNA and protein means $(\mathbb{E}\{R\}, \mathbb{E}\{P\})$:

$$\frac{d}{dt}\begin{bmatrix} \mathbb{E}\{R\} \\ \mathbb{E}\{P\} \end{bmatrix} = \begin{bmatrix} -\gamma_r & 0 \\ k_p & -\gamma_p \end{bmatrix} + \begin{bmatrix} k_r \\ 0 \end{bmatrix}.$$

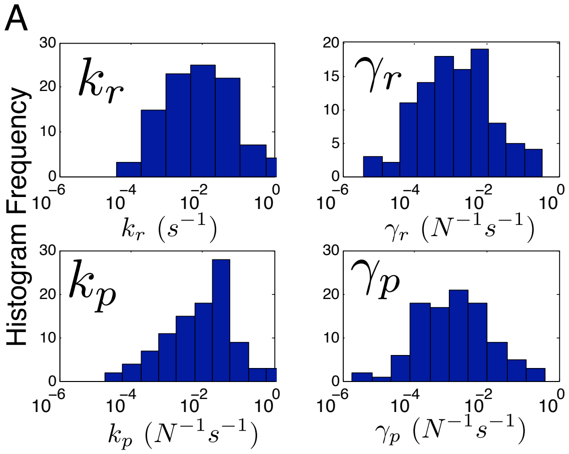
The solution for the mRNA and protein means at $t > 0$ can be written as:

$$\begin{aligned}
\mathbb{E}\{R(t)\} &= \mathrm{e}^{-\gamma_r t}\overline{R}_0 + \frac{k_r}{\gamma_r}\left(1 - \mathrm{e}^{-\gamma_r t}\right) \\
\mathbb{E}\{P(t)\} &= \mathrm{e}^{-\gamma_p t}\overline{P}_0 + \int_0^t \mathrm{e}^{-\gamma_p(t-\tau)} k_p \mathbb{E}\{R(\tau)\} d\tau \\
&= \mathrm{e}^{-\gamma_p t}\overline{P}_0 + \int_0^t \mathrm{e}^{-\gamma_p(t-\tau)} k_p \left(\mathrm{e}^{-\gamma_r \tau}\overline{R}_0 + \frac{k_r}{\gamma_r}\left(1 - \mathrm{e}^{-\gamma_r \tau}\right)\right) d\tau, \\
&= \mathrm{e}^{-\gamma_p t}\overline{P}_0 + k_p\overline{R}_0 \int_0^t \mathrm{e}^{-\gamma_p(t-\tau)}\mathrm{e}^{-\gamma_r \tau} d\tau + k_p k_r \int_0^t \mathrm{e}^{-\gamma_p(t-\tau)} \frac{1}{\gamma_r}\left(1 - \mathrm{e}^{-\gamma_r \tau}\right) d\tau,
\end{aligned}$$

where $\overline{R}_0$ and $\overline{P}_0$ are the initial expectations of the mRNA and proteins, respectively. By separating out all of the terms that depend upon unknowns $k_r$, $k_p$ and $\overline{R}_0$, the expression for the evolution of $\mathbb{E}\{P(t)\}$ can be rewritten as:

$$\mathbb{E}\{P(t)\} = f_1(\gamma_1, t)\overline{P}_0 + f_2(\gamma_p, \gamma_r, t)k_p\overline{R}_0 + f_3(\gamma_p, \gamma_r, t)k_p k_r.$$

Suppose that the parameters $\gamma_r$ and $\gamma_p$ are known, and that $\mathbb{E}\{P\}$ can be measured at any point in time. In this case, it is immediately obvious that *at best* one can only determine the two products $k_p\overline{R}_0$ and $k_p k_r$, but one cannot individually determine any of the three individual parameters $k_r$, $k_p$, or $\overline{R}_0$ without some additional piece of information, such as measurements describing the variation in the protein populations.



B $\mathbb{E}\{R\}$, $\mathbb{E}\{R^2\}$, $\mathbb{E}\{P\}$, $\mathbb{E}\{P^2\}$, $\mathbb{E}\{RP\}$

| Number of Time Points \ Number of Experiments | 1 | 2 | 4 |
|---|---|---|---|
| 10 | 13 | 31 | 46 |
| 20 | 26 | 54 | 73 |
| 40 | 46 | 69 | 81 |

C $\mathbb{E}\{R\}$, $\mathbb{E}\{P\}$

| Number of Time Points \ Number of Experiments | 1 | 2 | 4 |
|---|---|---|---|
| 10 | 2 | 12 | 27 |
| 20 | 15 | 33 | 45 |
| 40 | 26 | 48 | 67 |

D $\mathbb{E}\{P\}$, $\mathbb{E}\{P^2\}$

| Number of Time Points \ Number of Experiments | 1 | 2 | 4 |
|---|---|---|---|
| 10 | 1 | 7 | 11 |
| 20 | 4 | 12 | 29 |
| 40 | 9 | 31 | 47 |

Figure 5: Different strategies to identify transcription and translation parameters from simulated data with *known* initial conditions. (A) One hundred different models have been generated, each with log-normally distributed parameters. Histograms of these parameters are shown and cover about four different orders of magnitude. Each parameter set and initial distribution at time $t = 0$ defines a trajectory of the moments. These trajectories are assumed to be fully (B) or partially (C,D) measured at various points in time, but with unbiased additive measurement errors of $\pm 10\%$. (B-D) An estimation strategy is considered to be successful if it manages to identify all four parameters $\{k_r, \gamma_r, k_p, \gamma_p\}$ each within an error of $\pm 5\%$, and each table lists the number of successes per one hundred systems: (B) with full measurement of the moments: $(\mathbb{E}[R], \mathbb{E}[P], \mathbb{E}[R^2], \mathbb{E}[P^2], \mathbb{E}[RP])$, (C) with measurement of only the means: $(\mathbb{E}[R], \mathbb{E}[P])$, and (D) with measurement of only the protein marginals: $(\mathbb{E}[P], \mathbb{E}[P^2])$. For each strategy, the identification is done by using one, two or four different experiments each with a different *known* initial condition. Also, different numbers of measurement points are considered which include 10, 20 or 40 points in time, each separated by a time of 100 seconds.

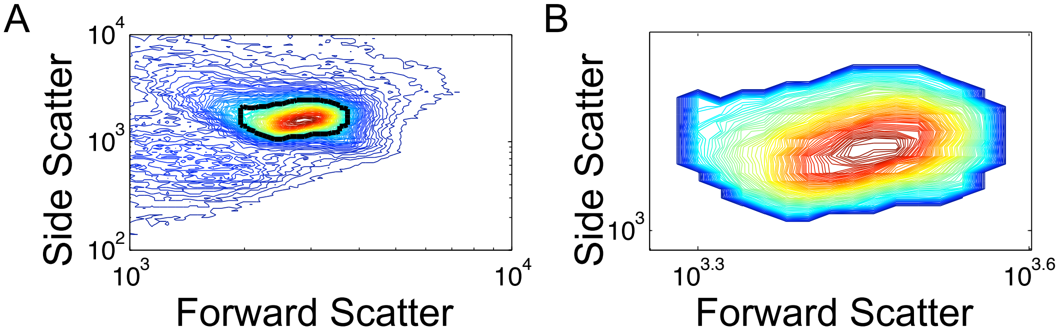


Figure 6: Gating of flow cytometry measurements based upon forward and side scatter data. A) Histogram of the forward and side scatter measurements of *gfp* without IPTG induction before gating. The black line represents the selected set of cells for which the histogram is one third its maximal height or greater. B) After gating only those cells within the densely populated gating region are kept for later analysis.